MULTISCALE MODEL. SIMUL.
Vol. 18, No. 2, pp. 1153–1177

RANDOMIZED SAMPLING FOR BASIS FUNCTION CONSTRUCTION IN GENERALIZED FINITE ELEMENT METHODS*

KE CHEN†, QIN LI‡, JIANFENG LU§, AND STEPHEN J. WRIGHT¶

Abstract. In the framework of generalized finite element methods for elliptic equations with rough coefficients, efficiency and accuracy of the numerical method depend critically on the use of appropriate basis functions. This work explores several random sampling strategies that construct approximations to the optimal set of basis functions of a given dimension, and proposes a quantitative criterion to analyze and compare these sampling strategies. Numerical evidence shows that the best results are achieved by two strategies, Random Gaussian and Smooth Boundary sampling.

Key words. randomized sampling, generalized finite element methods, random singular value decomposition, elliptic equation, generalized eigenvalue problem

AMS subject classification. 65N30

DOI. 10.1137/18M1166432

1. Introduction.

This paper considers techniques for constructing basis functions for generalized finite element methods applied to elliptic equations with rough coefficients. The elliptic partial differential equation is

$$\begin{cases} -\operatorname{div}(a(x)\nabla u(x)) = f(x), & x \in \Omega, \\ u(x) = 0, & x \in \partial\Omega, \end{cases} \tag{1.1}$$

with $f \in L^2(\Omega)$ and a uniformly elliptic coefficient function $a \in L^\infty(\Omega)$, that is, there exist $\alpha_*, \beta_* > 0$ such that $a(x) \in [\alpha_*, \beta_*]$ for all $x \in \Omega$. Note that we assume only L^∞ regularity of a, so the coefficient could be rather rough, which poses challenges for conventional numerical methods, such as the standard finite element method with local polynomial basis functions.

Numerical methods can be designed to take advantage of certain analytical properties of problem (1.1). A classical example is when a is two-scale, that is, $a(x) = a_0(x, \frac{x}{\varepsilon})$, where $a_0(x, y)$ is 1-periodic with respect to its second argument. (Thus, ε characterizes explicitly the small scale of the problem.) Using the theory of homogenization [4, 24], several numerical methods have been proposed over the past few decades to capture the homogenized solution of the problem and possibly also to pro-

*Received by the editors January 22, 2018; accepted for publication (in revised form) August 23, 2019; published electronically June 25, 2020.
https://doi.org/10.1137/18M1166432
Funding: The second author was supported in part by the National Science Foundation under grants DMS-1619778 and DMS-1750488. The third author was supported in part by the National Science Foundation under grant DMS-1454939. The fourth author was supported in part by NSF awards IIS-1447449, 1628384, and 1634597 and AFOSR award FA9550-13-1-0138. The first, second, and fourth authors were supported by TRIPODS: NSF 1740707.
†Department of Mathematics, University of Texas at Austin, Austin, TX 78731 (kechen@math.utexas.edu).
‡Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706 (qinli@math.wisc.edu).
§Department of Mathematics, Department of Physics, and Department of Chemistry, Duke University, Durham, NC 27708 (jianfeng@math.duke.edu).
¶Computer Sciences Department, University of Wisconsin-Madison, Madison, WI 53706 (swright@cs.wisc.edu).

vide some microscopic information. Approaches of this type include the multiscale finite element method [8, 13, 14, 12] and the heterogeneous multiscale method [6, 7, 20].

While methods designed for numerical homogenization can be applied to the cases of rough media ($a \in L^\infty$), the lack of favorable structural properties often degrades the efficiency and convergence rates. Various numerical methods have been proposed for L^∞ media, including the generalized finite element method [2], upscaling based on harmonic coordinates [22], elliptic solvers based on $\mathcal{H}$-matrices [3, 10], and Bayesian numerical homogenization [21], to name just a few. Our work uses the framework of the generalized finite element method (gFEM) of [2]. The idea is to approximate the local solution space by constructing good basis functions locally and to use either the partition-of-unity or the discontinuous Galerkin method to obtain a global discretization.

According to the partition-of-unity finite element theory, which we will recall briefly in section 2, the global error is controlled by the accuracy of the numerical local solution spaces. Thus, global performance of the method depends critically on efficient preparation of accurate local solution spaces. Towards this end, Babuška and Lipton [1] studied the Kolmogorov width of a finite-dimensional approximation to the optimal solution space, and showed that the Kolmogorov width decays almost exponentially fast, as we will recall in section 2. Different deterministic approaches under this framework with oversampling are investigated and numerically compared in [26]. The basis construction algorithm proposed in [1] follows the analysis closely: a full list of a-harmonic functions (up to discretization) is obtained in each patch, and local basis functions are obtained by a "postprocessing" step of solving a generalized eigenvalue problem to select modes with highest "energy ratios." Since the roughness of a necessitates a fine discretization in each patch, and thus a large number of a-harmonic functions per patch, the overall computational cost of this strategy to construct local basis functions is high.

Our work is based on the gFEM framework [2] together with the concept of optimal local solution space via Kolmogorov width studied in [1]. The idea of introducing random sampling or oversampling to construct local basis functions is studied in [5, 9, 17], where these methods are shown to be computationally effective. However, a systematic investigation of random sampling in the context of numerical PDEs is lacking. There is no criterion that justifies the "goodness" of basis functions constructed through random sampling. The main contribution of this paper is twofold. We systematically examine these random sampling approaches and introduce a criterion that evaluates different sets of basis functions; furthermore, we propose a random projection method that obtains a set of a-harmonic basis functions automatically. Randomized algorithms have been shown to be powerful in reducing computational complexities in looking for low rank factorization of matrices. Since the generalized eigenvalues decay almost exponentially, the local solution space is of approximate low rank, and random sampling approaches can capture this space effectively. The efficiency of the approach certainly depends on the particular random sampling strategy employed; we explore several strategies and identify the most successful ones.

As mentioned above, the idea of random sampling or oversampling to construct local basis functions is not completely new. In [9], the authors proposed to compute a generalized eigenvalue problem (using the stiffness matrix and the mass matrix) as a postprocessing step for basis selection. Similar strategies have been considered in the discontinuous Galerkin framework [17], but these approaches require a full basis of local solutions. The random sampling strategy is incorporated in [5] to improve efficiency in the offline stage.

There are several important differences between our approaches and those of [9, 17, 5]. First, we provide a quantitative criterion for evaluating the efficiency and accuracy of different random sampling strategies. Second, we find that the best randomized sampling strategy is *not* necessarily based on randomly assigning boundary conditions, as done in [5]. As indicated by the proposed criterion, a good sampling strategy should eliminate boundary layers and maintain much of the "energies" of the samples in the interior. Third, instead of using a stiffness-mass ratio as done in [9], the selection process here is guided by the behavior of the restriction operator (see section 2), which is proved to be optimal in [1].

Other basis construction approaches based on the gFEM framework have been explored in the literature, mostly based on a similar offline-online strategy. In the offline step, one prepares the solution space (either local or global). In the online step, one assembles the basis through the Galerkin framework (see, for example, [19, 23, 15]). The random sampling strategy can also be explored in these contexts.

The organization of the rest of the paper is as follows. We review preliminaries in section 2, including the basics of basis construction and error analysis. In section 3, we describe the random sampling framework and present a few particular sampling strategies. We connect and compare our framework with the randomized singular value decomposition (rSVD) in section 3.3. To compare the various sampling approaches, according to the criterion we propose in section 4, a random sampling strategy with higher energies achieves smaller Kolmogorov distances to the optimal basis. Numerical examples in section 5 demonstrate the effectiveness of our approach.

This paper only serves as the first step towards evaluating randomly constructed basis functions, and there are many other choices and parameters that we do not fully investigate. One example is the ratio of the enlargement: a bigger enlarged domain gives faster decay in singular values, but the numerical cost is fairly high. These issues are left to future research.

2. Previous results and context. Here we provide some preliminary results about the gFEM, including the concept of low rank solution space, and review the construction of basis functions for the local solution space.

We restate the elliptic equation (1.1) as follows:

$$\begin{cases} \mathcal{L}u = -\operatorname{div}\left(a(x)\nabla u(x)\right) = f(x)\,, & x \in \Omega, \\ u(x) = 0\,, & x \in \partial\Omega, \end{cases} \tag{2.1}$$

with $0 < \alpha_* \le a(x) \le \beta_*$, where $\mathcal{L}$ denotes the elliptic operator. The weak formulation of (2.1) is

$$\langle a(x)\nabla u\,, \nabla v\rangle_{L^2(\Omega)} = \langle f\,, v\rangle_{L^2(\Omega)}$$

for all test functions v, where $\langle f, g\rangle_{L^2(\Omega)} := \int_\Omega f(x)g(x)\mathrm{d}x$.

In the Galerkin framework, one constructs the solution space first. Given the following approximation space, defined by basis functions ϕ_i, $i = 1, 2, \ldots, n$,

$$\operatorname{Span}\{\phi_i\,, i = 1, 2, \ldots, n\}\,, \tag{2.2}$$

we substitute the ansatz $U = \sum_{i=1}^n c_i\phi_i$ into (2.1) to obtain

$$\sum_{j=1}^{n} \langle a(x)\nabla\phi_j\,, \nabla\phi_i\rangle_{L^2(\Omega)} c_j = \langle f\,, \phi_i\rangle_{L^2(\Omega)}\,.$$

We write this system in matrix form as follows:

$$\mathsf{A}\vec{c} = \vec{b}, \tag{2.3}$$

where A is a symmetric matrix with entries $\mathsf{A}_{mn} = \langle a\nabla\phi_m\,, \nabla\phi_n\rangle_{L^2(\Omega)}$, and $\vec{c}$ (with $\vec{c}_m = c_m$) is a list of coefficients to be determined. The right-hand side is the load vector $\vec{b}$, with entries $\vec{b}_m = \langle f\,, \phi_m\rangle_{L^2(\Omega)}$.

It is well known that the following quasi-optimality holds:

$$\|u - U\|_{\mathcal{E}(\Omega)} \leq C\|u - \mathcal{P}u\|_{\mathcal{E}(\Omega)}\,,$$

where C is some constant depending on α_* and β_*, and $\mathcal{P}u$ is the projection of the true solution u onto the space (2.2). Here the energy norm on any subdomain $\omega \subset \Omega$ is defined by

$$\|v\|_{\mathcal{E}(\omega)} = \langle a\nabla v\,, \nabla v\rangle^{1/2}_{L^2(\omega)} := \left[\int_\omega a|\nabla v(x)|^2 \mathrm{d}x\right]^{1/2}. \tag{2.4}$$

Thus to guarantee small numerical error $\|u-U\|_{\mathcal{E}(\Omega)}$, we require a set of basis functions that form a space for which $\|u - \mathcal{P}u\|_{\mathcal{E}(\Omega)}$ is small.

The main difficulty of computing the elliptic equation with rough coefficient is that a large number of basis functions are apparently needed. When $a(x)$ is rough with ε as its smallest scale, for standard piecewise affine finite elements, the mesh size Δx needs to resolve the smallest scale, so that $\Delta x \ll \varepsilon$ in each dimension. It follows that the dimension n of the system (2.3) is $n = \mathcal{O}(1/\varepsilon^d) \gg 1$, where d is the spatial dimension. The large size of stiffness matrix A and its large condition number (usually on the order of $\mathcal{O}(1/\varepsilon^2)$) make the problem expensive to solve using this approach.

The question then is whether it is possible to design a Galerkin space for which n is independent of ε. As mentioned in section 1, the offline-online procedure makes this approach feasible, as we discuss next.

2.1. Generalized finite element method. The gFEM was one of the earliest methods to utilize the offline-online procedure. This approach is based on the partition of unity. One first decomposes the domain Ω into many small patches ω_i, $i = 1, 2, \ldots, m$, that form an open cover of Ω. Each patch ω_i is assigned a partition-of-unity function ν_i that is zero outside ω_i and 1 over most of the set ω_i. Specifically, there is a positive constant C such that

$$0 \leq \nu_i(x) \leq 1 \qquad \text{for all } x \in \Omega \text{ and all } i = 1, 2, \ldots, m, \tag{2.5a}$$

$$\nu_i(x) = 0 \qquad \text{for all } x \in \Omega \setminus \omega_i \text{ and all } i = 1, 2, \ldots, m, \tag{2.5b}$$

$$\max_{x\in\Omega} |\nabla\nu_i(x)| \leq \frac{C}{\operatorname{diam}(\omega_i)} \qquad \text{for all } i = 1, 2, \ldots, m. \tag{2.5c}$$

Moreover, we have

$$\sum_{i=1}^{m} \nu_i(x) = 1 \quad \text{for all } x \in \Omega\,. \tag{2.6}$$

In the offline step, basis functions $\phi_{i,j}$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n_i$, are constructed for each patch ω_i, where n_i is the number of basis functions in patch i. We denote the numerical local solution space in patch ω_i by

$$\Phi_{[i]} = \operatorname{Span}\{\phi_{i,j},\, j = 1, 2, \ldots, n_i\}\,. \tag{2.7}$$

In the online step, the Galerkin formulation is used, with the space in (2.2) replaced by

$$\Phi := \bigoplus_{i=1,2,\ldots,m} \Phi_{[i]}\nu_i = \mathrm{Span}\{\phi_{i,j}\nu_i,\ i=1,2,\ldots,m,\ j=1,2,\ldots,n_i\}\,. \tag{2.8}$$

Details can be found in [2].

The total number of basis functions is $\sum_{i=1}^m n_i$. If all n_i, $i=1,2,\ldots,m$, are bounded by a modest constant, the dimension of the space Φ is of order m, so the computation in the online step is potentially inexpensive. It is proved in [2] that the total approximation error is governed by the sum of all local approximation errors.

THEOREM 2.1. *Denote by u the solution to* (2.1). *Suppose $\{\omega_i\}_{i=1,2,\ldots,m}$ forms an open cover of Ω, and let $\{\nu_i\}_{i=1,2,\ldots,m}$ denote the set of partition-of-unity functions defined in* (2.5). *If the solution can be approximated well by $\zeta_i \in \Phi_{[i]}$ in each patch ω_i, the global error is small too. Specifically, if we assume that*

$$\|u-\zeta_i\|_{L^2(\omega_i)} \le \varepsilon_1(i) \quad \textit{and} \quad \|u-\zeta_i\|_{\mathcal{E}(\omega_i)} \le \varepsilon_2(i)\,, \quad i=1,2,\ldots,m, \tag{2.9}$$

and define

$$\zeta(x) = \sum_{i=1}^m \zeta_i(x)\nu_i(x)\,,$$

then $\zeta(x) \in H^1(\Omega)$, and for the constant C defined in (2.5), *we have*

$$\|u-\zeta\|_{L^2(\Omega)} \le \max_i \|\nu_i\|_\infty \left(\sum_{i=1}^m \varepsilon_1^2(i)\right)^{1/2}$$

and

$$\|u-\zeta\|_{\mathcal{E}(\Omega)} \le C\left(\sum_{i=1}^m \frac{\varepsilon_1^2(i)}{\mathrm{diam}^2(\omega_i)} + \max_i \|\nu_i\|_\infty^2 \sum_{i=1}^m \varepsilon_2^2(i)\right)^{1/2}.$$

This theorem shows that the approximation error of the Galerkin numerical solution for the gFEM depends directly on the accuracy of the local approximation spaces in each patch.

2.2. Low rank local solution space. One reason for the success of gFEM is that the local numerical solution space is approximately low rank, meaning that n_i has a modest value for all i in (2.7); see [1]. We review the relevant results in this section, and show how to find $\Phi_{[i]}$.

Denote by ω_i^* an enlargement of the patch ω_i, that is, a set for which $\omega_i \subset \omega_i^* \subset \Omega$. To simplify notation, we suppress subscripts i from here on. We introduce a restriction operator:

$$P : H_a(\omega^*)/\mathbb{R} \to H_a(\omega)/\mathbb{R}\,,$$

where $H_a(\omega^*)$ is the collection of all a-harmonic functions in ω^* and $H_a(\omega^*)/\mathbb{R}$ represents the quotient space of $H_a(\omega^*)$ with respect to the constant function. (This modification is needed to make $\|\cdot\|_{\mathcal{E}(\omega^*)}$ a norm, since an a-harmonic function is defined only up to an additive constant.) The operator P is determined uniquely by $a(x)$ restricted in ω^* and ω. We denote its adjoint operator by $P^* : H_a(\omega)/\mathbb{R} \to H_a(\omega^*)/\mathbb{R}$. It is shown in [1] that the operator P^*P is a compact, self-adjoint, nonnegative operator on $H_a(\omega^*)/\mathbb{R}$.

To derive an n-dimensional approximation of $H_a(\omega^*)/\mathbb{R}$, we define as follows the *Kolmogorov distance* of an arbitrary n-dimensional function subspace $S_n \subset H_a(\omega)/\mathbb{R}$ to $H_a(\omega^*)/\mathbb{R}$ associated with their corresponding norms $\|\cdot\|_{\mathcal{E}(\omega)}$ and $\|\cdot\|_{\mathcal{E}(\omega^*)}$, respectively:

$$d(S_n, H_a(\omega^*)) = \sup_{\substack{u \in H_a(\omega^*)/\mathbb{R}, \\ \|u\|_{\mathcal{E}(\omega^*)} \le 1}} \inf_{\xi \in S_n} \|Pu - \xi\|_{\mathcal{E}(\omega)} . \tag{2.10}$$

(We omit the norms from the arguments of d, since they are clear from the context.) By considering all possible S_n, we can identify the optimal approximation space Φ_n that achieves the infimum:

$$\Phi_n := \arg\inf_{S_n} d(S_n, H_a(\omega^*)) . \tag{2.11}$$

We now define a distance measure between ω and ω^* as follows:

$$d_n(\omega, \omega^*) = d(\Phi_n, H_a(\omega^*)) .$$

The term $d_n(\omega, \omega^*)$ is the celebrated Kolmogorov n-width of the compact operator P. It reflects how quickly a-harmonic functions supported on ω^* lose their energies when confined to ω. According to [25], the optimal approximation space Φ_n and Kolmogorov n-width can be found explicitly, in terms of the eigendecomposition of P^*P on ω^*, which is

$$P^*P\psi_i = \lambda_i \psi_i , \quad i = 1, 2, \ldots , \tag{2.12}$$

with λ_i arranged in descending order and $\{\psi_i,\ i = 1, 2, \ldots\}$ the corresponding eigenvectors, which are automatically orthonormal according to $\langle \cdot, \cdot \rangle_{\mathcal{E}(\omega^*)}$. By defining

$$\Psi_n := \mathrm{Span}\{\psi_1, \ldots , \psi_n\}, \tag{2.13}$$

the optimal approximation space is

$$\Phi_n := P\Psi_n = \mathrm{Span}\{\phi_1, \phi_2, \ldots , \phi_n\} , \quad \text{with } \phi_i := P\psi_i,\ i = 1, 2, \ldots , n. \tag{2.14}$$

It follows from the definitions above that

$$d_n(\omega, \omega^*) = \sqrt{\lambda_{n+1}} . \tag{2.15}$$

Note that ψ_i are all supported in the enlarged domain ω^*, while ϕ_i are their confinements in ω. Almost-exponential decay of the Kolmogorov width with respect to n was proved in [1, Theorem 3.3], according to the following result.

THEOREM 2.2. *The accuracy $d_n(\omega, \omega^*)$ has nearly exponential decay for n sufficiently large: for any small $\varepsilon > 0$, we have*

$$d_n(\omega, \omega^*) \le e^{-n^{(d+1)^{-1}-\varepsilon}} .$$

It follows that for any function u that is an a-harmonic function in the patch $\omega^* \subset \mathbb{R}^2$, we can find a function $v \in \Phi_n$ for which

$$\|u - v\|_{\mathcal{E}(\omega)} \le d_n(\omega, \omega^*) \|u\|_{\mathcal{E}(\omega^*)} \lesssim e^{-n^{1/3}-\varepsilon} .$$

Remark 1. Note that d_n is the $(n+1)$th singular value $\sqrt{\lambda_{n+1}}$ of P. Because of the fast decay of d_n with respect to n indicated by Theorem 2.2, P is an approximately low rank operator. It follows that almost all a-harmonic functions supported on ω^*, when confined in ω, look almost alike, and can be represented by a relatively small number of "important" modes.

Remark 2. We note that enlarging the domain for oversampling is a standard approach: in [12], the boundary layer behavior confined in ω^*/ω was studied and utilized for computation.

2.3. Computing the local solution space. We describe here the computation of an approximation to Φ_n via discretized versions of the objects defined in the previous subsection. More specifically, we discretize the enlarged patch ω^* with a fine mesh, and collect all a-harmonic functions upon discretization. To collect all a-harmonic functions, we would need to solve the system with elliptic operator (1.1) locally, with all possible Dirichlet boundary conditions on $\partial\omega^*$. For ease of presentation, here and in what follows, we assume that we choose a piecewise-affine finite element discretization of the patch for computing the local a-harmonic functions. Then the discretized a-harmonic functions are determined by their values on grid points $\{y_1, y_2, \ldots, y_{N_y}\}$ on the boundary of $\partial\omega^*$. We proceed in three stages.

Stage A. Construct the discrete a-harmonic function space $H_a(\omega^*)$ on the fine mesh via the functions χ_i obtained by solving the following system for $i = 1, 2, \ldots, N_y$:

$$\begin{cases} \mathcal{L}\chi_i = -\mathrm{div}\,(a(x)\nabla\chi_i) = 0\,, & x \in \omega^*, \\ \chi|_{\partial\omega^*} = \delta_i\,, & y_i \in \partial\omega^*, \end{cases} \tag{2.16}$$

where δ_i is the hat function that peaks at y_i and equals zero at other grid points y_j, $j \neq i$. Recall that we have assumed a piecewise-affine finite element discretization of ω^*.

Stage B. Compute the eigenvalue problem (2.12) in the space spanned by $\{\chi_i\,, i = 1, 2, \ldots, N_y\}$. Noting that

$$\langle P^*P\psi_i\,, \delta\rangle_{\mathcal{E}(\omega^*)} = \langle P\psi_i\,, P\delta\rangle_{\mathcal{E}(\omega)} = \langle \psi_i\,, \delta\rangle_{\mathcal{E}(\omega)} \quad \text{for all } \delta \in \mathrm{Span}\{\chi_1, \ldots, \chi_{N_y}\}\,, \tag{2.17}$$

the weak formulation of the eigenvalue problem (2.12), when confined in the discrete a-harmonic function space, is given by

$$\langle \psi_i\,, \chi\rangle_{\mathcal{E}(\omega)} = \lambda_i \langle \psi_i\,, \chi\rangle_{\mathcal{E}(\omega^*)} \quad \text{for all } \chi \in \mathrm{Span}\{\chi_1, \ldots, \chi_{N_y}\}\,.$$

Expanding the eigenfunction ψ_i in terms of χ_j, $j = 1, 2, \ldots, N_y$, as

$$\psi_i = \sum_j c_j^{(i)} \chi_j, \tag{2.18}$$

we obtain the following equation for the coefficient vector $\vec{c}^{(i)}$:

$$\sum_j c_j^{(i)} \langle \chi_j\,, \chi_k\rangle_{\mathcal{E}(\omega)} = \lambda_i \sum_j c_j^{(i)} \langle \chi_j\,, \chi_k\rangle_{\mathcal{E}(\omega^*)}\,.$$

This system can be written as a generalized eigenvalue problem, as follows:

$$\mathsf{S}\vec{c}^{(i)} = \lambda_i \mathsf{S}^* \vec{c}^{(i)}\,, \tag{2.19}$$

with $\mathsf{S}_{mn} = \langle \chi_m\,, \chi_n \rangle_{\mathcal{E}(\omega)}$ and $\mathsf{S}^*_{mn} = \langle \chi_m\,, \chi_n \rangle_{\mathcal{E}(\omega^*)},\ m, n = 1, 2, \ldots, N_y$.

This generalized eigenvalue problem can be solved for $\lambda_i,\ i = 1, 2, \ldots, N_y$, arranged in descending order, and their associated eigenfunctions $\psi_i,\ i = 1, 2, \ldots, N_y$, defined from (2.18) using the generalized eigenvectors $\vec{c}^{(i)}$ from (2.19). Choose index n to satisfy $\lambda_{n+1} < \mathrm{TOL} < \lambda_n$, where TOL is a given error tolerance.

Stage C. Obtain Φ_n by substituting the functions $\psi_i,\ i = 1, 2, \ldots, n$, calculated in Stage B, into (2.14).

3. Randomized sampling methods for local bases. In this section we propose a class of random sampling methods to construct local basis functions efficiently. As seen in section 2.3, finding the optimal basis functions amounts to solving the generalized eigenvalue problem in (2.19). The main cost comes not from performing the eigenvalue decomposition, but rather from computing the a-harmonic functions χ_i, which are used to construct the matrices S and S^* in (2.19). As shown in section 2.2, the eigenvalues decay almost exponentially, indicating that only a limited number of local modes is needed to represent the whole solution space well. This low rank structure motivates us to consider randomized sampling techniques.

Randomized algorithms have been highly successful in compressed sensing, where they are used to extract low rank structure efficiently from data. The Johnson–Lindenstrauss lemma [16] suggests that structure in high-dimensional data points is largely preserved when projected onto random lower-dimensional spaces. The rSVD algorithm uses this idea to capture the principal components of a large matrix by random projection of its row and column spaces onto smaller subspaces; see [11] for a review. In the current numerical PDE context, knowing that the local solution space is essentially low rank, we seek to adopt the random sampling idea to generate local approximate solution spaces efficiently.

Randomized SVD cannot be applied directly in our context, as we discuss in section 3.3. We propose instead a method based on Galerkin approximation of the generalized eigenvalue problem on a small subspace. One immediate difficulty is that an arbitrarily given random function is not necessarily a-harmonic. Thus, our method first generates a random collection of functions and projects them onto the a-harmonic function space, and then solves the generalized eigenvalue problem (2.19) on the subspace to find the optimal basis functions. A detailed description of our approach is shown in Algorithm 1.

Note that the steps in Stage 2 of Algorithm 1 are parallel to those of section 2.3, but only a small number of functions γ_k are used in the generalized eigenvalue problem, rather than the whole list of a-harmonic functions (i.e., $N_r \ll N_y$). We therefore save significant computation in preparing the a-harmonic function space, in assembling the S and S^* matrices, and in solving the generalized eigenvalue decomposition.

The key is to use the random sampling strategy in Stage 1 of Algorithm 1 to generate an effective small subspace for the generalized eigenvalue problem. This aspect of the algorithm will be the focus of the rest of this section.

3.1. a-harmonic projection. Let us first discuss the a-harmonic projection of a given function ξ supported on ω^*. This problem can be formulated as a PDE-

Algorithm 1. Determining optimal local bases.

Stage 1: Randomly generate a collection of N_r a-harmonic functions.

Stage 1-A: Randomly pick functions $\{\xi_k : k = 1, 2, \ldots, N_r\}$ supported on ω^*.

Stage 1-B: For each $k = 1, 2, \ldots, N_r$, project ξ_k onto the a-harmonic function space to obtain γ_k.

Stage 2: Solve the generalized eigenvalue problem to determine leading modes.

Stage 2-A: Define

$$\mathsf{S}_{\gamma,mn} := \langle \gamma_m \, , \gamma_n \rangle_{\mathcal{E}(\omega)} \, , \quad \mathsf{S}^*_{\gamma,mn} := \langle \gamma_m \, , \gamma_n \rangle_{\mathcal{E}(\omega^*)} \, , \quad m, n = 1, 2, \ldots, N_r, \tag{3.1}$$

and solve the associated generalized eigenvalue problem

$$\mathsf{S}_\gamma \vec{v}^\gamma = \lambda^\gamma \mathsf{S}^*_\gamma \vec{v}^\gamma \, , \tag{3.2}$$

with $(\lambda^\gamma_j \, , \vec{v}^\gamma_j)$ denoting the jth eigenpairs, such that $\lambda^\gamma_1 \geq \lambda^\gamma_2 \geq \cdots \geq \lambda^\gamma_{N_r} \geq 0$.

Stage 2-B: Choose n such that $\lambda^\gamma_1 \geq \cdots \geq \lambda^\gamma_n \geq \mathrm{TOL} > \lambda^\gamma_{n+1}$ (where TOL is a preset tolerance), and collect the first n eigenfunctions to use as the local basis functions:

$$\Phi^r_n = \mathrm{Span}\{\phi^r_1, \phi^r_2, \ldots, \phi^r_n\} = \mathrm{Span}\{P\psi^r_j, \, j = 1, 2, \ldots, n\} \, , \tag{3.3}$$

where $\psi^r_j = \sum_k \vec{v}^\gamma_{j,k} \gamma_k$.

constrained optimization problem:

$$\min_\gamma \frac{1}{2} \|\gamma - \xi\|^2_{L^2(\omega^*)} \quad \text{subject to} \quad \mathcal{L}\gamma = 0 \, , \tag{3.4}$$

where $\mathcal{L} = -\operatorname{div} a\nabla$ is the elliptic operator defined in (2.1). The Lagrangian function for (3.4) is as follows:

$$F(\gamma, \mu) := \frac{1}{2} \|\gamma - \xi\|^2_{L^2(\omega^*)} - \langle \mu \, , \mathcal{L}\gamma \rangle_{L^2(\omega^*)} \, , \tag{3.5}$$

where μ is a Lagrange multiplier. In the discrete setting, we form a grid $\{x_i\}$ over ω^* and denote by ζ_i the hat function centered at grid point x_i. (Recall that we have assumed piecewise-affine finite element discretization.) The Lagrangian function for the corresponding discretized optimization problem is

$$F(\gamma, \mu) = \frac{1}{2} (\gamma^i - \xi^i)^\top (\gamma^i - \xi^i) - \mu^\top \mathsf{A}^{ii} \gamma^i - \mu^\top \mathsf{A}^{ib} \gamma^b \, , \tag{3.6}$$

where the superindices i and b stand for interior and boundary grids, respectively, and A is the stiffness matrix whose (m, n) element is

$$\mathsf{A}_{mn} = \langle a\nabla\zeta_m \, , \nabla\zeta_n \rangle_{L^2(\omega^*)} \, .$$

In the discrete setting, μ is a vector of the same length as γ^i (the number of grid points in the interior). Note that in the translation to the discrete setting, we represent $\mathcal{L}\gamma = 0$ by $\mathsf{A}\gamma = 0$, which leads to

$$\mathsf{A}^{ii} \gamma^i + \mathsf{A}^{ib} \gamma^b = 0 \, .$$

Here A^{ii} is the stiffness matrix confined in the interior, and A^{ib} is the part of the stiffness matrix generated by taking the inner product of the interior basis functions and the boundary basis functions. To solve the minimization problem, we take the partial derivatives of (3.6) with respect to γ and μ and set them equal to zero, as follows:

$$\begin{aligned}\nabla_{\gamma^i} F &= \gamma^i - \xi^i - \mathsf{A}^{ii\top}\mu = 0\,,\\ \nabla_{\gamma^b} F &= \mathsf{A}^{ib\top}\mu \qquad\qquad = 0\,,\\ \nabla_{\mu} F &= \mathsf{A}^{ii}\gamma^i + \mathsf{A}^{ib}\gamma^b \quad = 0\,.\end{aligned}$$

Some manipulation yields the following systems for γ^b and γ^i:

$$\mathsf{A}^{ib\top}\left(\mathsf{A}^{ii}\right)^{-2}\mathsf{A}^{ib}\gamma^b = -\mathsf{A}^{ib\top}\left(\mathsf{A}^{ii}\right)^{-1}\xi^i, \quad \gamma^i = -\left(\mathsf{A}^{ii}\right)^{-1}\mathsf{A}^{ib}\gamma^b\,.$$

The solution to this system gives the solution of (3.4) in the discrete setting. Recall that γ^b is a vector containing only the boundary conditions for the solution, and thus the computation is rather cheap, given that the matrix $\mathsf{A}^{ib\top}\left(\mathsf{A}^{ii}\right)^{-2}\mathsf{A}^{ib}$ can be prepared ahead of time. Computing γ^i using γ^b amounts to numerically solving a finite element problem confined in a small domain ω^*, and thus the numerical cost is the same as preparing an a-harmonic function.

3.2. Random sampling strategies. We have many possible choices for the random functions ξ_k, $k = 1, 2, \ldots, N_r$, in Stage 1-A of Algorithm 1. Here we list several natural choices.

1. *Interior δ-function.* Choose a random grid point in ω, and set $\xi(x) = 1$ at this grid point and zero at all other grid points. That is, ξ is the hat function associated with the grid point x.
2. *Interior independent and identically distributed (i.i.d.) function.* Choose the value of ξ at each grid point in ω independently from a standard normal Gaussian distribution. The values of ξ at grid points in $\omega^*\backslash\omega$ are set to 0.
3. *Full-domain i.i.d. function.* This choice is the same as in strategy 2, except that the values of ξ at the grid points in $\omega^*\backslash\omega$ are also chosen as Gaussian random variables.
4. *Random Gaussian.* Choose a random grid point $x_0 \in \omega$, and set $\xi(x) = e^{-\frac{(x-x_0)^2}{2}}$ at all grid points $x \in \omega^*$.

We aim to select basis functions (through Stage 2) that are associated with the largest eigenvalues, so that the Kolmogorov n-width can be small (2.15). Thus, we hope that in Stage 1, the chosen functions ξ_k provide large eigenvalues λ_i in (2.19). A large value of λ indicates that a large portion of the energy is maintained in ω, with only a small amount coming from the buffer region $\omega^*\backslash\omega$. It therefore suggests choosing functions ξ_k with most of their variations inside ω. However, the projection onto an a-harmonic space step makes the locality of the resulting functions hard to predict. In section 4, we propose and analyze a criterion for the performance of the random sampling schemes. In particular, we compare the four choices listed above.

We mention here that a list of a-harmonic functions could be obtained through a different route: one can prepare boundary conditions and compute local a-harmonic functions inside ω^* with the preassigned boundary. There are various ways to prepare boundary conditions, including the following.

5. *Random i.i.d. boundary sampling.* In [5], the authors proposed to obtain a list of random a-harmonic functions by computing the local elliptic equation

with i.i.d. random Dirichlet boundary conditions. Assuming there are N_y grid points on the boundary $\partial\omega^*$, we define g to be a vector of length N_y with i.i.d. random variables for each component. We then define γ by solving

$$\begin{cases} \mathcal{L}\gamma = 0\,, & x \in \omega^*, \\ \gamma|_{\partial\omega^*} = g. \end{cases} \tag{3.7}$$

This process is repeated N_r times to obtain a set of N_r random a-harmonic functions $\{\gamma_k : k = 1, 2, \ldots, N_r\}$.

6. *Randomized boundary sampling with exponential covariance.* A technique in which the Dirichlet boundary conditions are chosen to be random Gaussian variables with a specified covariance matrix is described in [18]. This matrix is assumed to be an exponential function, that is,

$$\mathrm{Cov}(y_i, y_j) = \exp(-|y_i - y_j|/\sigma)\,. \tag{3.8}$$

The first few modes of a Karhunen–Loéve expansion are used to construct a boundary condition in (3.7), with which basis functions are computed. Although a justification for this approach is not provided, numerical computations show that it is more efficient than the i.i.d. random boundary sampling.

7. *Smooth Boundary sampling.* Since i.i.d. random Dirichlet boundary conditions typically yield solutions that oscillate a lot near the boundary, they have sharp boundary layers. To eliminate this effect, one can use a Gaussian kernel to smooth out the boundary profile. In particular, the i.i.d. random sample can be convolved with a Gaussian function $\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$ to obtain a smoother boundary condition.

We note that strategies 5 and 6 above were proposed in [5] and [18], respectively. However, in [5], the postprocessing for basis selection was conducted using the generalized eigenvalue problem of the stiffness and mass matrix instead of (2.12), and thus there is no guarantee in the exponential decay.

3.3. Connection with randomized SVD. We briefly address the connection between the random sampling method we propose in this paper and the well-known randomized SVD (rSVD) algorithm. Although rSVD cannot be used directly in our problem, it serves as a motivation for our randomized sampling strategies.

The rSVD algorithm, studied thoroughly in [11], speeds up the computation of the SVD of a matrix when the matrix is large and approximately low rank. With high probability, the singular vector structure is largely preserved when the matrix is projected onto a random subspace. Specifically, for a random matrix R with a small number of columns (the number depending on the rank of A), it is proved in [11] that if we obtain Q from the QR factorization of AR, we have

$$\|\mathsf{A} - \mathsf{Q}\mathsf{Q}^\top \mathsf{A}\|_2 \ll \|\mathsf{A}\|_2\,. \tag{3.9}$$

This bound implies that any vector in the range space of A can be well approximated by its projection onto the space spanned by Q. For example, if $\vec{u} = \mathsf{A}\vec{v}$, we have from (3.9) that

$$\|\vec{u} - \mathsf{Q}\mathsf{Q}^\top \vec{u}\| \ll \|\vec{u}\|\,. \tag{3.10}$$

We note that Q and AR span the same column space, but Q is easier to work with and better conditioned, because its columns are orthonormal. Equivalent to (3.10),

we can also say that any $\vec{u}$ in the image of A can be approximated well using a linear combination of the columns of AR.

To see the connection between rSVD and our problem, we first write the generalized eigenvalue problem (2.19) in an SVD form. Recall the definitions (3.1) of S and S^*,

$$\mathsf{S}_{mn} = \int_\omega a(x)\nabla\chi_m(x)\cdot\nabla\chi_n(x)\mathrm{d}x\,, \quad \mathsf{S}^*_{mn} = \int_{\omega^*} a(x)\nabla\chi_m(x)\cdot\nabla\chi_n(x)\mathrm{d}x\,,$$

and define

$$\Phi^* = \left[\sqrt{a}\nabla\chi_1, \sqrt{a}\nabla\chi_2, \ldots, \sqrt{a}\nabla\chi_{N_y}\right], \quad \Phi = \Phi^*|_\omega\,. \tag{3.11}$$

Since $\mathsf{S} = \Phi^\top\Phi$ and $\mathsf{S}^* = \Phi^{*\top}\Phi^*$, the generalized eigenvalue problem (2.19) can be written as follows:

$$\Phi^\top\Phi\vec{c} = \lambda\Phi^{*\top}\Phi^*\vec{c}. \tag{3.12}$$

We write the QR factorization for Φ^* as

$$\Phi^* = \mathsf{Q}_{\Phi^*}\mathsf{R}_{\Phi^*}\,,$$

and denote $\vec{d} = \mathsf{R}_{\Phi^*}\vec{c}$. By substituting into (3.12), we obtain

$$\left(\Phi\mathsf{R}_{\Phi^*}^{-1}\right)^\top\left(\Phi\mathsf{R}_{\Phi^*}^{-1}\right)\vec{d} = \lambda\vec{d},$$

meaning that $(\sqrt{\lambda}, \vec{d})$ forms a singular value pair of the matrix $\Phi R_{\Phi^*}^{-1}$.

According to the rSVD argument, the leading singular vectors of $\Phi\mathsf{R}_{\Phi^*}^{-1}$ are captured by those of

$$\Phi\mathsf{R}_{\Phi^*}^{-1}\mathsf{R}\,, \tag{3.13}$$

where R is a matrix whose entries are i.i.d. Gaussian random variables. Specifically, with high probability, the leading singular values of $\Phi\mathsf{R}_{\Phi^*}^{-1}\mathsf{R}$ are almost the same as those of $\Phi\mathsf{R}_{\Phi^*}^{-1}$, and the column space spanned by (3.13) largely covers the image of $\Phi\mathsf{R}_{\Phi^*}^{-1}$, as in (3.11).

We now interpret $\Phi\mathsf{R}_{\Phi^*}^{-1}\mathsf{R}$ from the viewpoint of PDEs. Decomposing $\mathsf{R}_{\Phi^*}^{-1}\mathsf{R}$ into columns as

$$\mathsf{R}_{\Phi^*}^{-1}\mathsf{R} = [r_1, r_2, \ldots]\,, \quad \text{with} \quad r_k = [r_{k1}, r_{k2}, \ldots]^\top\,, \tag{3.14}$$

and denoting $\gamma_k = \sum_j r_{kj}\chi_j$, we have from (3.11) that

$$\Phi r_k = \sqrt{a}\nabla\left(\sum_j r_{kj}\chi_j\right) \doteq \sqrt{a}\nabla\gamma_k\,.$$

Numerically, this corresponds to solving the following system for γ_k:

$$\begin{cases}\mathcal{L}\gamma_k = -\mathrm{div}\,(a(x)\nabla\gamma_k) = 0\,, \quad x\in\omega^*,\\ \gamma_k|_{\partial\omega^*} = \sum_j r_{kj}\delta_{y_j}.\end{cases} \tag{3.15}$$

It is apparent from this equation that to obtain $\Phi\mathsf{R}_{\Phi^*}^{-1}\mathsf{R}$, we do not need to compute all functions χ_j, $j = 1, 2, \ldots, N_y$, and use them to construct γ_k. Rather, we can compute

γ_k directly by solving the elliptic equation with random boundary conditions given by r_{kj}, $j = 1, 2, \ldots, N_y$. The cost of this procedure is proportional to N_r, which is much less than N_y.

Unfortunately, this procedure is difficult to implement in a manner that accords with the rSVD theory. R is constructed using i.i.d. Gaussian random variables, but $\mathsf{R}_{\Phi^*}^{-1}$ is unknown ahead of time, so the distribution of r_k defined in (3.14) is unknown. The theory here suggests that there exists some random sampling strategy that achieves the accuracy and efficiency that characterize rSVD, but it does not provide such a strategy.

4. Efficiency of various random sampling methods. As discussed in section 3, given multiple ways to choose the random samples in Stage 1 of Algorithm 1, it is natural to ask which one is better, and how to predetermine the approximation accuracy. We answer these questions in this section.

The key requirement is that Algorithm 1 should capture the high-energy modes of (2.12), the modes that correspond to the highest values of λ_i. We start with a simple example in section 4.1 that finds the relationship between the energy captured by a certain single mode, and the angle that this mode makes with the highest energy mode. The argument used can be easily applied to the case with multiple modes, and the link towards the Kolmogorov distance will be shown in section 4.2. We will discuss the situation in the general setting with plain linear algebra, and its relevance to local PDE basis construction is outlined in section 4.3.

4.1. A one-mode example. Suppose we are working in a three-dimensional space, with symmetric positive definite matrices A and B and generalized eigenvectors x_1, x_2, and x_3 such that

$$\langle x_i\,, x_j\rangle_{\mathsf{B}} = x_i^\top \mathsf{B} x_i = \delta_{ij}\,, \quad \langle x_i\,, x_j\rangle_{\mathsf{A}} = x_i^\top \mathsf{A} x_i = \delta_{ij}\lambda_i \tag{4.1}$$

for generalized eigenvalues $\lambda_1 > \lambda_2 > \lambda_3$. We thus have

$$\mathsf{A}x_i = \lambda_i \mathsf{B}x_i, \quad i = 1, 2, 3.$$

Suppose we have some one-dimensional space $\mathcal{X}$ spanned by a vector x, and we intend to use it as an approximation of the space $\mathcal{X}_1$ spanned by the leading eigenvector x_1. The energy of $\mathcal{X}$ is

$$E(\mathcal{X}) = \frac{x^\top \mathsf{A} x}{x^\top \mathsf{B} x}\,, \tag{4.2}$$

and the angle between the spaces $\mathcal{X}$ and $\mathcal{X}_1$ is defined by

$$d(\mathcal{X}, \mathcal{X}_1) = \max_{|\beta|\le 1} \min_{\alpha} \|\alpha x - \beta x_1\|_{\mathsf{A}}\,. \tag{4.3}$$

We have the following result (which generalizes easily to dimension greater than 3).

PROPOSITION 1. *The angle* (4.3) *is bounded in terms of the energy* (4.2) *as follows:*

$$d(\mathcal{X}, \mathcal{X}_1) \le \sqrt{\frac{\lambda_1\lambda_2\,(\lambda_1 - E(\mathcal{X}))}{(\lambda_1 - \lambda_2)E(\mathcal{X})}}\,. \tag{4.4}$$

Proof. The proof is simple algebra. As $\{x_1, x_2, x_3\}$ span the entire space and are B-orthogonal, we have

$$x = w_1 x_1 + w_2 x_2 + w_3 x_3\,, \tag{4.5}$$

with $w_i = x^\top \mathsf{B} x_i$, $i = 1, 2, 3$. According to the definition of the angle, one can reduce the problem by setting $\beta = 1$ and $\sum_i w_i^2 = 1$, so that $\|x\|_{\mathsf{B}} = 1$ in (4.3). (With these normalizations, we have from (4.1) and (4.2) that $E(\mathcal{X}) = x^T A x = \lambda_1 w_1^2 + \lambda_2 w_2^2 + \lambda_3 w_3^2$.) We thus have

$$\begin{aligned} d(\mathcal{X}, \mathcal{X}_1)^2 &= \min_\alpha \|\alpha x - x_1\|_{\mathsf{A}}^2 \\ &= \min_\alpha \|(\alpha w_1 - 1)x_1 + \alpha w_2 x_2 + \alpha w_3 x_3\|_{\mathsf{A}}^2 \\ &= \min_\alpha \left((\alpha w_1 - 1)^2 \lambda_1 + \alpha^2 w_2^2 \lambda_2 + \alpha^2 w_3^2 \lambda_3\right). \end{aligned}$$

The minimum is achieved at $\alpha = w_1\lambda_1 / E(\mathcal{X})$, with the minimized angle being

$$\left(\angle(x, x_1)\right)^2 = \frac{E(\mathcal{X}) - w_1^2 \lambda_1}{E(\mathcal{X})} \lambda_1\,. \tag{4.6}$$

To bound the numerator in (4.6) we observe that

$$E(\mathcal{X}) - w_1^2\lambda_1 = w_2^2\lambda_2 + w_3^2\lambda_3 \le w_2^2\lambda_2 + w_3^2\lambda_2 = (1 - w_1^2)\lambda_2,$$

and moreover,

$$E(\mathcal{X}) \le w_1^2\lambda_1 + (1-w_1^2)\lambda_2 \;\Rightarrow\; \lambda_1 - E(\mathcal{X}) \ge (1-w_1^2)(\lambda_1 - \lambda_2) \;\Rightarrow\; 1 - w_1^2 \le \frac{\lambda_1 - E(\mathcal{X})}{\lambda_1 - \lambda_2}\,.$$

By combining these last two bounds, we obtain

$$E(\mathcal{X}) - w_1^2\lambda_1 \le \lambda_2 \frac{\lambda_1 - E(\mathcal{X})}{\lambda_1 - \lambda_2}.$$

By substituting this bound into (4.6), we obtain (4.4). ∎

Note that the bound (4.4) decreases to zero as $\lambda_1 - E(\mathcal{X}) \downarrow 0$.

According to (4.4), a larger gap in the spectrum between λ_1 and λ_2 yields a tighter bound, thus better control over the angle. The theorem indicates that the "energy" is the quantity that measures how well the randomly given vector x captures the first mode, and thus serves as the criterion for the quality of the approximation.

4.2. Higher-dimensional criteria. In this section, we seek the counterpart in higher-dimensional space of the previous result. Suppose now that the two symmetric positive definite matrices A and B are $n \times n$, and their generalized eigenpairs (λ_i, x_i) satisfy the conditions

$$\langle x_i\,, x_j \rangle_{\mathsf{B}} = \delta_{ij}\,, \quad \langle x_i\,, x_j \rangle_{\mathsf{A}} = \delta_{ij}\lambda_i\,, \tag{4.7}$$

so that

$$\mathsf{A}x_i = \lambda_i \mathsf{B} x_i\,, \quad \text{with} \quad \lambda_1 \ge \cdots \ge \lambda_k > \lambda_{k+1} \ge \cdots \ge \lambda_n > 0\,,$$

that is,

$$\mathsf{A}\mathsf{X} = \mathsf{B}\mathsf{X}\Lambda\,, \quad \text{with} \quad \Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)\,. \tag{4.8}$$

Suppose we are trying to recover the optimal k-dimensional space $\mathcal{X}^h := \mathrm{Span}\{\mathsf{X}^h\}$, where $\mathsf{X}^h = [x_1, x_2, \ldots, x_k]$ collects the first k eigenfunctions. Define $\mathcal{X}^l := \mathrm{Span}\{\mathsf{X}^l\}$, where $\mathsf{X}^l = [x_{k+1}, \ldots, x_n]$ collects the remaining modes. Denoting by $\mathcal{Y}$ our proposed approximation space to $\mathcal{X}^h$, we seek a quantity that measures how well the proposal space $\mathcal{Y}$ approximates the optimal space $\mathcal{X}^h$. In particular, we will show below that the "angle" between the proposal $\mathcal{Y}$ and the to-be-recovered space $\mathcal{X}^h$ relies on the "energy" of $\mathcal{Y}$.

DEFINITION 4.1 (energy of a space $\mathcal{Z}$). *For any given k-dimensional space $\mathcal{Z}$, define $\mathsf{Z} \in \mathbb{R}^{n\times k}$ to be a matrix whose columns form a B-orthonormal basis of $\mathcal{Z}$ (obtained through performing Gram–Schmidt orthogonalization with B-inner product). Then the energy of $\mathcal{Z}$ is defined as*

$$E(\mathcal{Z}) := \frac{\mathrm{Tr}(\mathsf{Z}^\top \mathsf{A} \mathsf{Z})}{\mathrm{Tr}(\mathsf{Z}^\top \mathsf{B} \mathsf{Z})}. \tag{4.9}$$

We note that this is a natural extension of energy defined in (4.2), and it is well defined in the sense that the energy term (4.9) depends solely on the space $\mathcal{Z}$ rather than on the basis Z, as shown in Appendix B.

We now generalize the angle (4.3) and define the Kolmogorov distance from space $\mathcal{Y}$ to the optimal space $\mathcal{X}^h$, with norms $\|\cdot\|_{\mathsf{A}}$ and $\|\cdot\|_{\mathsf{B}}$, respectively.

DEFINITION 4.2 (angle between spaces). *Define the Kolmogorov distance from $\mathcal{Y}$ to the optimal subspace $\mathcal{X}^h$ as follows:*

$$d(\mathcal{Y}, \mathcal{X}^h) = \sup_{\substack{z\in\mathcal{X}^h,\\ \|z\|_{\mathsf{B}}\le 1}} \inf_{y\in\mathcal{Y}} \|z - y\|_{\mathsf{A}}. \tag{4.10}$$

Notice that $d(\mathcal{Y}, \mathcal{X}^h)$ is a discrete version of (2.10), so we don't have operator P in (4.10) since it is implicit in the norm $\|\cdot\|_{\mathsf{A}}$.

Similarly to the previous section, we show that $E(\mathcal{Y})$ is related to $d(\mathcal{Y}, \mathcal{X}^h)$. In Definition 4.1 the energy $E(\mathcal{Z})$ is defined for a B-orthonormal basis Z, and for consistency we assume that $\mathsf{Y} \in \mathbb{R}^{n\times k}$ collects a B-orthonormal basis of space $\mathcal{Y}$.

Since X spans the entire space, we can express Y as follows:

$$\mathsf{Y} = \mathsf{X}\mathsf{C} = \mathsf{X}^h\mathsf{C}^h + \mathsf{X}^l\mathsf{C}^l, \tag{4.11}$$

where $\mathsf{C} \in \mathbb{R}^{n\times k}$. The columns of C are orthonormal because from (4.7) and the definition of Y, we have

$$\mathsf{Y} = \mathsf{X}\mathsf{C} \;\Rightarrow\; \mathbb{1} = \mathsf{Y}^\top\mathsf{B}\mathsf{Y} = \mathsf{C}^\top\mathsf{X}^\top\mathsf{B}\mathsf{X}\mathsf{C} = \mathsf{C}^\top\mathsf{C}. \tag{4.12}$$

We denote by C^h the upper $\mathbb{R}^{k\times k}$ portion of C, and by C^l the lower $\mathbb{R}^{(n-k)\times k}$ portion. Denoting the elements of C by c_{ji}, we have

$$\mathsf{C}^h = [c_{ji}]_{j=1,2,\ldots,k;\, i=1,2,\ldots,k}, \quad \mathsf{C}^l = [c_{ji}]_{j=k+1,k+2,\ldots,n;\, i=1,2,\ldots,k}. \tag{4.13}$$

By orthonormality of C, it follows that

$$\sum_{j=1}^{k} c_{ji}^2 + \sum_{j=k+1}^{n} c_{ji}^2 = 1, \quad i = 1, 2, \ldots, k,$$

and thus

$$\left[C^{l\top}C^l\right]_{ii} = 1 - \left[C^{h\top}C^h\right]_{ii} = 1 - \sum_{j=1}^{k} c_{ji}^2, \quad i = 1, 2, \ldots, k. \tag{4.14}$$

LEMMA 4.1. *The trace of* $\mathsf{C}^{l\top}\mathsf{C}^l$ *is bounded by the energy difference between the optimal space* $\mathcal{X}^h$ *and the proposed space* $\mathcal{Y}$*:*

$$\operatorname{Tr}(\mathsf{C}^{l\top}\mathsf{C}^l) \leq \frac{k\left(E(\mathcal{X}^h) - E(\mathcal{Y})\right)}{\lambda_k - \lambda_{k+1}}. \tag{4.15}$$

Furthermore, C^h *is invertible if*

$$E(\mathcal{X}^h) - E(\mathcal{Y}) < \frac{\lambda_k - \lambda_{k+1}}{k}. \tag{4.16}$$

Proof. We have from (4.12) that

$$\mathsf{C}^{h\top}\mathsf{C}^h + \mathsf{C}^{l\top}\mathsf{C}^l = \mathbb{1}. \tag{4.17}$$

Since both X^h and Y are B-orthonormal and have k columns, we have

$$\operatorname{Tr}\left(\mathsf{X}^{h\top}\mathsf{B}\mathsf{X}^h\right) = \operatorname{Tr}\left(\mathsf{Y}^\top\mathsf{B}\mathsf{Y}\right) = k.$$

By substituting X^h and Y into the definition of energy (4.9), we have

$$k\left(E(\mathcal{X}^h) - E(\mathcal{Y})\right) = \operatorname{Tr}\left(\mathsf{X}^{h\top}\mathsf{A}\mathsf{X}^h - \mathsf{Y}^\top\mathsf{A}\mathsf{Y}\right).$$

By substituting for Y from (4.11), and using (4.7), we have

$$\begin{aligned} k\left(E(\mathcal{X}^h) - E(\mathcal{Y})\right) &= \operatorname{Tr}\left(\mathsf{X}^{h\top}\mathsf{A}\mathsf{X}^h - \mathsf{C}^{h\top}\mathsf{X}^{h\top}\mathsf{A}\mathsf{X}^h\mathsf{C}^h - \mathsf{C}^{l\top}\mathsf{X}^{l\top}\mathsf{A}\mathsf{X}^l\mathsf{C}^l\right) \\ &= \operatorname{Tr}\left(\Lambda^h - \mathsf{C}^{h\top}\Lambda^h\mathsf{C}^h - \mathsf{C}^{l\top}\Lambda^l\mathsf{C}^l\right), \end{aligned} \tag{4.18}$$

where $\Lambda^h := \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$ and $\Lambda^l := \operatorname{diag}(\lambda_{k+1}, \lambda_{k+2}, \ldots, \lambda_n)$. For the terms on the right-hand side of (4.18), we have that

$$\operatorname{Tr}\left(\mathsf{C}^{l\top}\Lambda^l\mathsf{C}^l\right) \leq \lambda_{k+1}\operatorname{Tr}\left(\mathsf{C}^{l\top}\mathsf{C}^l\right), \tag{4.19}$$

and that

$$\begin{aligned} \operatorname{Tr}\left(\Lambda^h - \mathsf{C}^{h\top}\Lambda^h\mathsf{C}^h\right) &= \sum_{j=1}^k \lambda_j - \sum_{i=1}^k\sum_{j=1}^k \lambda_j c_{ji}^2 \\ &= \sum_{j=1}^k \lambda_j\left(1 - \sum_{i=1}^k c_{ji}^2\right) \\ &\geq \lambda_k \sum_{j=1}^k \left(1 - \sum_{i=1}^k c_{ji}^2\right) \\ &= \lambda_k \operatorname{Tr}\left(\mathsf{C}^{l\top}\mathsf{C}^l\right), \end{aligned} \tag{4.20}$$

where we used (4.14). By substituting (4.19) and (4.20) into (4.18), we obtain

$$k\left(E(\mathcal{X}^h) - E(\mathcal{Y})\right) \geq (\lambda_k - \lambda_{k+1})\operatorname{Tr}\left(\mathsf{C}^{l\top}\mathsf{C}^l\right),$$

which is equivalent to (4.15).

When condition (4.16) holds, we have from (4.15) that $\operatorname{Tr}(\mathsf{C}^{l\top}\mathsf{C}^l) < 1$. Thus since $\mathsf{C}^{h\top}\mathsf{C}^h = \mathbb{1} - \mathsf{C}^{l\top}\mathsf{C}^l$, and setting $\|\mathsf{C}^{l\top}\mathsf{C}^l\| \leq \operatorname{Tr}(\mathsf{C}^{l\top}\mathsf{C}^l) < 1$, we have that $\mathsf{C}^{h\top}\mathsf{C}^h$ is nonsingular, so that the $k \times k$ matrix C^h is nonsingular. ∎

We finally use energy distance $E(\mathcal{X}^h)-E(\mathcal{Y})$ to estimate the Kolmogorov distance $d(\mathcal{Y},\mathcal{X}^h)$ as follows.

THEOREM 4.1. *Considering the optimal space $\mathcal{X}^h$ and the proposed space $\mathcal{Y}$, if*

$$E(\mathcal{X}^h) - E(\mathcal{Y}) \leq \frac{\lambda_k - \lambda_{k+1}}{2k}\,, \tag{4.21}$$

then we have

$$d(\mathcal{Y},\mathcal{X}^h) \leq \sqrt{\lambda_{k+1}\frac{\|\mathsf{C}^{l\top}\mathsf{C}^l\|}{1-\|\mathsf{C}^{l\top}\mathsf{C}^l\|}}\,, \tag{4.22}$$

and furthermore,

$$d(\mathcal{Y},\mathcal{X}^h) \leq \sqrt{2\lambda_{k+1}\frac{k\,(E(\mathcal{X}^h)-E(\mathcal{Y}))}{\lambda_k-\lambda_{k+1}}}\,. \tag{4.23}$$

Proof. Choosing an arbitrary $z = \mathsf{X}^h\alpha$ with $\|\alpha\| \leq 1$, we look for β such that $y = \mathsf{Y}\beta$ is closest to z in the A-norm. The solution, obtained from the minimization problem

$$\min_\beta f_\alpha(\beta) := \|y-z\|_{\mathsf{A}}^2 = (\mathsf{Y}\beta - \mathsf{X}^h\alpha)^\top \mathsf{A}(\mathsf{Y}\beta - \mathsf{X}^h\alpha)\,, \tag{4.24}$$

is

$$\beta_\alpha^* = (\mathsf{Y}^\top\mathsf{A}\mathsf{Y})^{-1}\mathsf{Y}^\top\mathsf{A}\mathsf{X}^h\alpha\,. \tag{4.25}$$

Note from the definition (4.10) that

$$d(\mathcal{Y},\mathcal{X}^h) = \sup_{\|\alpha\|\leq 1}\sqrt{f_\alpha(\beta_\alpha^*)}. \tag{4.26}$$

From (4.7) and (4.11), we have

$$\mathsf{Y}^\top\mathsf{A}\mathsf{Y} = \mathsf{C}^\top\Lambda\mathsf{C} = \mathsf{C}^{h\top}\Lambda^h\mathsf{C}^h + \mathsf{C}^{l\top}\Lambda^l\mathsf{C}^l, \tag{4.27}$$

which is invertible, since C has orthonormal columns and Λ is diagonal and positive definite. Thus β_α^* is well defined by (4.25). By substituting (4.25) into (4.24), we obtain

$$f_\alpha(\beta_\alpha^*) = -\alpha^\top(\mathsf{Y}^\top\mathsf{A}\mathsf{X}^h)^\top(\mathsf{Y}^\top\mathsf{A}\mathsf{Y})^{-1}(\mathsf{Y}^\top\mathsf{A}\mathsf{X}^h)\alpha + \alpha^\top\mathsf{X}^{h\top}\mathsf{A}\mathsf{X}^h\alpha. \tag{4.28}$$

Note from (4.11) and (4.8) that

$$\mathsf{A}\mathsf{Y} = \mathsf{A}\mathsf{X}^h\mathsf{C}^h + \mathsf{A}\mathsf{X}^l\mathsf{C}^l = \mathsf{B}\mathsf{X}^h\Lambda^h\mathsf{C}^h + \mathsf{B}\mathsf{X}^l\Lambda^l\mathsf{C}^l,$$

so from (4.7), we have

$$\mathsf{X}^{h\top}\mathsf{A}\mathsf{Y} = (\mathsf{X}^{h\top}\mathsf{B}\mathsf{X}^h)\Lambda^h\mathsf{C}^h + (\mathsf{X}^{h\top}\mathsf{B}\mathsf{X}^l)\Lambda^l\mathsf{C}^l = \Lambda^h\mathsf{C}^h.$$

By substituting this equality together with (4.27) into (4.28), and using (4.7) again, we have

$$f_\alpha(\beta_\alpha^*) = -\alpha^\top(\Lambda^h\mathsf{C}^h)(\mathsf{C}^{h\top}\Lambda^h\mathsf{C}^h + \mathsf{C}^{l\top}\Lambda^l\mathsf{C}^l)^{-1}(\Lambda^h\mathsf{C}^h)^\top\alpha + \alpha^\top\Lambda^h\alpha\,. \tag{4.29}$$

Invertibility of C^h follows from Lemma 4.1 and the condition (4.21), so that $(\Lambda^h)^{1/2}\mathsf{C}^h$ is invertible, and we can transform (4.29) to

(4.30)
$$f_\alpha(\beta^*_\alpha) = -\alpha^\top(\Lambda^h)^{1/2}\left[\mathbb{1} + (\mathsf{C}^{h\top}(\Lambda^h)^{1/2})^{-1}\mathsf{C}^{l\top}\Lambda^l\mathsf{C}^l((\Lambda^h)^{1/2}\mathsf{C}^h)^{-1}\right]^{-1}(\Lambda^h)^{1/2}\alpha + \alpha^T\Lambda^h\alpha.$$

For any matrix A with $(\mathbb{1}+\mathsf{A})$ nonsingular, we have that

$$(\mathbb{1}+\mathsf{A})^{-1} = \mathbb{1} - \mathsf{A} + (\mathbb{1}+\mathsf{A})^{-1}\mathsf{A}^2. \tag{4.31}$$

Moreover, if A is symmetric positive semidefinite, then the last term $(\mathbb{1}+\mathsf{A})^{-1}\mathsf{A}^2$ is symmetric positive semidefinite, since if we write the eigenvalue decomposition of A as $\mathsf{A} = \mathsf{U}\mathsf{S}\mathsf{U}^\top$, where U is orthogonal and S is nonnegative diagonal, we have that $(\mathbb{1}+\mathsf{A})^{-1}\mathsf{A}^2 = \mathsf{A}^2(\mathbb{1}+\mathsf{A})^{-1} = \mathsf{U}(\mathbb{1}+\mathsf{S})^{-1}\mathsf{S}^2\mathsf{U}^\top$. Thus for any vector z, we have from (4.31) that

$$-z^\top(\mathbb{1}+\mathsf{A})^{-1}z + z^Tz \le -z^T(\mathbb{1}-\mathsf{A})z + z^Tz = z^T\mathsf{A}z.$$

By substituting $\mathsf{A} = (\mathsf{C}^{h\top}(\Lambda^h)^{1/2})^{-1}\mathsf{C}^{l\top}\Lambda^l\mathsf{C}^l((\Lambda^h)^{1/2}\mathsf{C}^h)^{-1}$ and $z = (\Lambda^h)^{1/2}\alpha$ into this expression, we have from (4.30) that

$$\begin{aligned} f_\alpha(\beta^*_\alpha) &\le \alpha^\top(\Lambda^h)^{1/2}(\mathsf{C}^{h\top}(\Lambda^h)^{1/2})^{-1}\mathsf{C}^{l\top}\Lambda^l\mathsf{C}^l((\Lambda^h)^{1/2}\mathsf{C}^h)^{-1}(\Lambda^h)^{1/2}\alpha \\ &= \alpha^\top(\mathsf{C}^h)^{-\top}\mathsf{C}^{l\top}\Lambda^l\mathsf{C}^l(\mathsf{C}^h)^{-1}\alpha \\ &\le \|\alpha\|^2\|(\mathsf{C}^{h\top}\mathsf{C}^h)^{-1}\|\|\mathsf{C}^{l\top}\Lambda^l\mathsf{C}^l\|. \end{aligned} \tag{4.32}$$

Note that $\|\alpha\| \le 1$, $\|\mathsf{C}^{l\top}\Lambda^l\mathsf{C}^l\| \le \lambda_{k+1}\|\mathsf{C}^{l\top}\mathsf{C}^l\|$, and

$$\left\|(\mathsf{C}^{h\top}\mathsf{C}^h)^{-1}\right\| = \left\|(\mathbb{1}-\mathsf{C}^{l\top}\mathsf{C}^l)^{-1}\right\| \le \sum_{i=0}^{\infty}\left\|\mathsf{C}^{l\top}\mathsf{C}^l\right\|^i = \frac{1}{1-\|\mathsf{C}^{l\top}\mathsf{C}\|},$$

so by substituting into (4.32), we have

$$f_\alpha(\beta^*_\alpha) \le \lambda_{k+1}\frac{\|\mathsf{C}^{l\top}\mathsf{C}^l\|}{1-\|\mathsf{C}^{l\top}\mathsf{C}^l\|} \quad \text{for all } \alpha \text{ with } \|\alpha\| \le 1, \tag{4.33}$$

which because of (4.26) yields (4.22).

Under condition (4.21) we have from Lemma 4.1 that

$$\|\mathsf{C}^{l\top}\mathsf{C}\| \le \mathrm{Tr}(\mathsf{C}^{l\top}\mathsf{C}) \le \frac{k\left(E(\mathcal{X}^h) - E(\mathcal{Y})\right)}{\lambda_k - \lambda_{k+1}} \le \frac{1}{2},$$

so that

$$\frac{\|\mathsf{C}^{l\top}\mathsf{C}^l\|}{1-\|\mathsf{C}^{l\top}\mathsf{C}^l\|} \le 2\|\mathsf{C}^{l\top}\mathsf{C}^l\| \le 2\frac{k\left(E(\mathcal{X}^h) - E(\mathcal{Y})\right)}{\lambda_k - \lambda_{k+1}},$$

yielding (4.23). ∎

4.3. Criteria used in random sampling for local basis functions. In our local basis construction problem, we identify A and B in section 4.2 with S and S^*, respectively, from (2.19). An energy term is constructed similarly.

DEFINITION 4.3 (energy of a function space). *Given the function space Γ_n, let $\{\tilde\gamma_1, \tilde\gamma_2, \ldots, \tilde\gamma_n\}$ be an $\mathcal{E}(\omega^*)$-orthonormal basis for Γ_n. The energy of Γ_n is defined by*

$$E(\Gamma_n) := \frac{\sum_{i=1}^n \langle a(x)|\nabla_x\tilde\gamma_i|^2\rangle_{\mathcal{E}(\omega)}}{\sum_{i=1}^n \langle a(x)|\nabla_x\tilde\gamma_i|^2\rangle_{\mathcal{E}(\omega^*)}}. \tag{4.34}$$

According to Theorem 4.1, a larger value of E indicates a smaller angle to the optimal basis set, and thus a better sampling strategy.

Theorem 4.1 suggests that a sampling strategy that provides a matrix Y of discretized basis functions with higher energy $E(\mathcal{Y})$ (closer to the optimal value of $E(\mathcal{X}^h)$) will result in a smaller Kolmogorov distance, and thus a better approximation to the optimal space $\mathcal{X}^h$. Larger values of E are achieved when the samples have their energies largely supported in the interior. This further suggests that construction of a-harmonic functions through random sampling of singular boundary conditions may not be the best strategy, because the boundary layer close to ω^* quickly damps out the solution and the energies concentrated in the margin $\omega^*\backslash\omega$, leading to relatively small energy in the interior and a smaller value of E in (4.34). These observations are borne out by the numerical experiments reported in the next section. Sampling strategies that avoid boundary layers are thereby preferred, which suggests that Random Gaussian (strategy 4) and random Smooth Boundary sampling (strategy 7) are likely to give better results. Our computational results support this claim.

We note that enlarging oversampling size is another efficient way of getting rid of boundary layers, but that generally leads to a smaller energy value in (4.34).

THEOREM 4.2. *In a two-dimensional space, for any small $\varepsilon > 0$ and n sufficiently large, given any $u \in H_a(\omega^*)/\mathbb{R}$ and a subspace $\Gamma_n = \mathrm{Span}\{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ spanned by random samples of a-harmonic functions, the accuracy of approximating u with a function γ from Γ_n is bounded by the following estimate:*

$$\min_{\gamma\in\Gamma_n} \|u-\gamma\|_{\mathcal{E}(\omega)} \leq \|u\|_{\mathcal{E}(\omega^*)} \left(e^{-n^{1/3-\varepsilon}} + d(\Gamma_n, \Psi_n) \right) ,$$

where Ψ_n is defined in (2.13) *and $d(\Gamma_n, \Psi_n)$ is defined in* (4.10).

Proof. Without loss of generality, we assume $\|u\|_{\mathcal{E}(\omega^*)} = 1$. Consider the optimal basis $\{\psi_i\}_{i=1}^{\infty} \subset H_a(\omega^*)/\mathbb{R}$ computed in (2.12), for which we have

$$\langle \psi_i, \psi_j \rangle_{\mathcal{E}(\omega^*)} = \delta_{ij} , \quad \langle \psi_i, \psi_j \rangle_{\mathcal{E}(\omega)} = \lambda_i \delta_{ij}.$$

We therefore have scalars $u_1, u_2, \ldots$ such that

$$u = \sum_{i=1}^{\infty} u_i \psi_i, \quad \sum_{i=1}^{\infty} u_i^2 = 1 .$$

Defining $\tilde{u} \in H_a(\omega^*)/\mathbb{R}$ by

$$\tilde{u} = \sum_{i=1}^{n} u_i \psi_i ,$$

the restriction $v = P\tilde{u} \in H_a(\omega)/\mathbb{R}$ has

$$\|u - v\|_{\mathcal{E}(\omega)} = \|u - P\tilde{u}\|_{\mathcal{E}(\omega)} = \left(\sum_{i=n+1}^{\infty} u_i^2 \lambda_i \right)^{1/2} .$$

By the definition (4.10) of Kolmogorov distance, there exists $\gamma \in \Gamma_n$ such that

$$\|v - \gamma\|_{\mathcal{E}(\omega)} \leq \|\tilde{u}\|_{\mathcal{E}(\omega^*)} d(\Gamma_n, \Psi_n) ,$$

where $\Psi_n := \mathrm{Span}\{\psi_1, \ldots, \psi_n\}$ as in (2.13). We further note that

$$\|\tilde{u}\|_{\mathcal{E}(\omega^*)} = \left(\sum_{i=1}^{n} u_i^2 \right)^{1/2} ,$$

and therefore

$$\begin{aligned}\|u-\gamma\|_{\mathcal{E}(\omega)} &\le \|u-v\|_{\mathcal{E}(\omega)} + \|v-\gamma\|_{\mathcal{E}(\omega)} \\ &\le \left(\sum_{i=n+1}^{\infty} u_i^2\lambda_i\right)^{1/2} + \left(\sum_{i=1}^{n} u_i^2\right)^{1/2} d(\Gamma_n,\Psi_n) \\ &\le (\lambda_{n+1})^{1/2} + d(\Gamma_n,\Psi_n) \\ &\le e^{-n^{1/3-\varepsilon}} + d(\Gamma_n,\Psi_n)\,,\end{aligned}$$

where the last inequality comes from Theorem 2.2. ∎

5. Computational results. We present numerical results in this section that show how the Kolmogorov distance of the random sampling subspace to the optimal space decreases with the number of basis functions, for different sampling strategies. Throughout this section, the domain ω and enlarged domain ω^* are defined by

$$\omega = [-1,1]\times[-1,1], \quad \omega^* = [-1.4,1.4]\times[-1.4,1.4].$$

The medium $a(x,y)$ is defined to be

$$\begin{aligned}a(x,y) = \frac{1}{5}\Bigg(&\frac{1.1+\sin(7\pi x)}{1.1+\sin(7\pi y)} + \frac{1.1+\sin(9\pi y)}{1.1+\cos(9\pi x)} + \frac{1.1+\cos(13\pi y)}{1.1+\cos(13\pi x)} \\ &+\frac{1.1+\cos(9\pi x)}{1.1+\sin(9\pi y)} + \frac{1.1+\sin(7\pi y)}{1.1+\sin(7\pi x)}\Bigg), \quad (x,y)\in\omega^*\,.\end{aligned} \tag{5.1}$$

Numerical results will be shown for discretization parameters $\mathrm{d}x = \mathrm{d}y = 1/40$.

The reference solution is obtained from the procedure summarized in section 2.3. To find the optimal solution space, we prepare the entire a-harmonic function space by going through all possible boundary condition configurations, before computing the general eigenvalue problem (2.19) for basis selection. This process requires computation of the elliptic equation (2.16) 444 times (each time with a hat function on the boundary of $\partial\omega^*$ concentrated at a specific grid point), followed by computation of the generalized eigenpairs of two matrices of size 444×444. We then implement all random sampling methods proposed in section 3. As we see below, the seven strategies have varying degrees of efficiency, but all capture the low rank structure of the optimal space.

High-energy modes. The first four modes $\{\phi_{1,2,3,4}\}$ of the reference solution are shown in Figure 5.1. These are obtained by following the procedure described in section 2.3. We note here the presence of boundary layers in ω^*, as the functions exhibit fine scale oscillations near the boundary $\partial\omega^*$; moreover, these oscillations in the boundary layer are trimmed away when the functions are confined to the patch ω.

Recovery of general eigenvalues. We now describe results obtained by random sampling methods with the seven sampling strategies. For each strategy, we sample only 20 a-harmonic functions for the computation in (3.2), hoping that these 20 random samples still capture the highest energy modes. In Figure 5.2, we plot (in log scale) the 20 generalized eigenvalues obtained from each of the seven sampling strategies, together with the leading 20 eigenvalues from the optimal reference solution. All methods give almost exponential decay of the eigenvalues. By far, Random Gaussian and Smooth Boundary (strategies 5 and 7) are the best two strategies for tracking the eigenvalues of the reference solution.

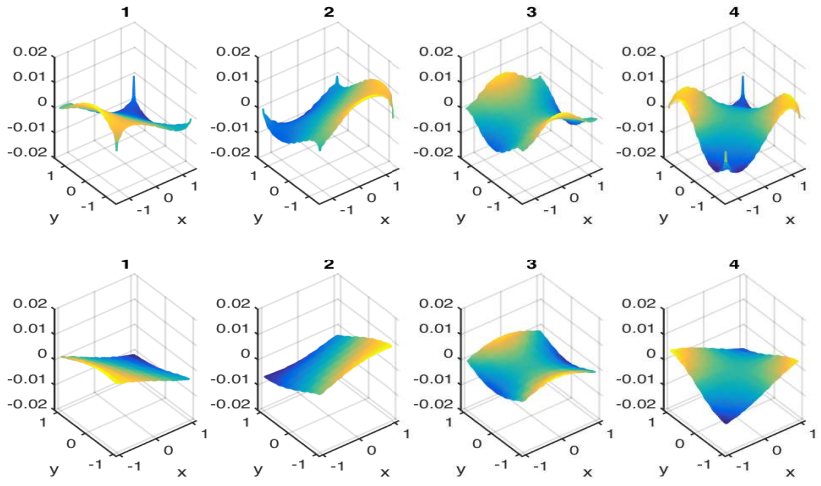

FIG. 5.1. *The first row shows the four modes $\phi_{1,2,3,4}$ supported on ω^*, and the second row shows the same modes confined in ω. Note that the boundary layers that appear in ω^* are not evident in ω.*

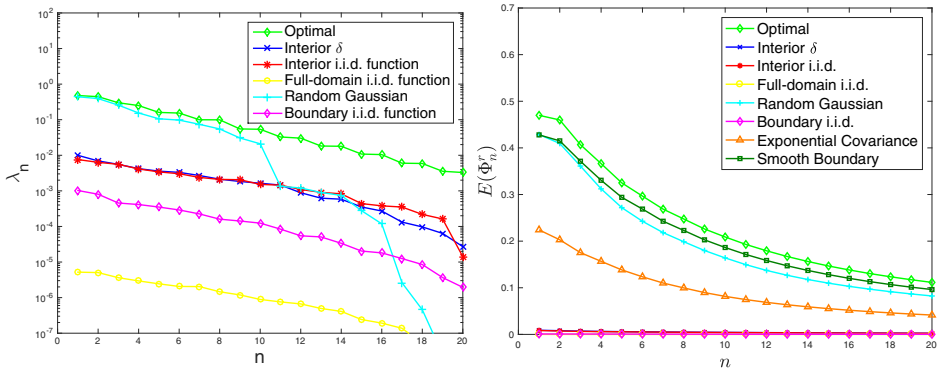

FIG. 5.2. Left: *Eigenvalues obtained from the six different random sampling strategies, using* 20 *samples each, and the leading* 20 *eigenvalues from the reference solution. Eigenvalues are computed from* (3.2) *for the random sampling strategies and from* (2.19) *for the reference solution. All methods show almost exponential decay of the eigenvalues, and Random Gaussian and Smooth Boundary are the best two sampling strategies in the sense that they match the reference eigenvalues most closely.* Right: *Energy $E(\Phi_n^r)$ of optimal space and approximate subspaces from different random sampling strategies using* 20 *samples. Energy is computed from* (4.34). *Again, the Random Gaussian and Smooth Boundary strategies achieve the minimal energy gap from the optimal space.*

It is expected that as the number of random samples increases, all random methods should do better at capturing the eigenvalues of the reference solution. This phenomenon is evident in Figure 5.3, where we use 300 random samples for all seven sampling strategies. All strategies except those involving the full-domain i.i.d. function and possibly the boundary i.i.d. function do well at matching the reference eigenvalues.

Figure 5.4 shows the recovery of eigenspace by random sampling procedures. We regard Φ_5, defined in (2.11), as the optimal space (the space expanded by the five modes with highest energies), and use Φ_m^r defined in (3.3) to approximate it for

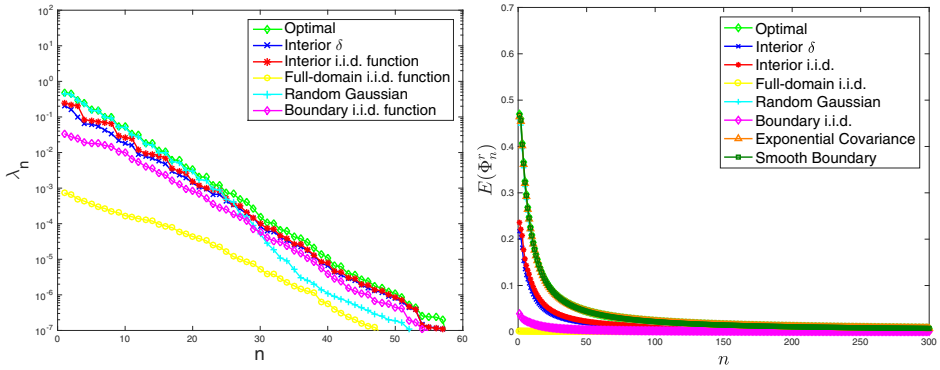

FIG. 5.3. *The same as Figure* 5.2, *but with* 300 *random samples instead of* 20. *Since* $N_r \sim N_y$, *the eigenvalues and energies obtained from random sampling tend to match the reference eigenvalues and optimal energy more closely.*

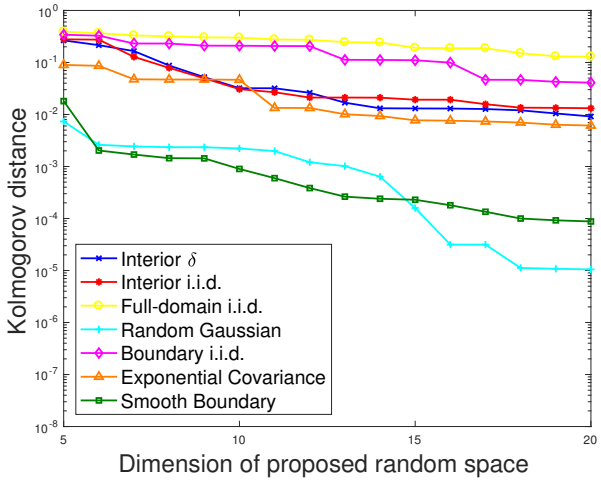

FIG. 5.4. *Kolmogorov distance of* Φ_5 *to* Φ_m^r *for* $m = 5, 6, \ldots, 20$ *sampling modes.*

$m = 5, 6, \ldots, 20$. The vertical axis shows Kolmogorov distance, whose computation is described in Appendix A. As expected, using more random modes leads to better recovery and thus smaller Kolmogorov distance (4.10). The plots show Kolmogorov distance decays roughly exponentially fast with respect to m, for all five sampling strategies. Once again, the Random Gaussian and Smooth Boundary strategies are by far the best: Φ_{20}^r approximates Φ_5 with accuracy near 10^{-4} or 10^{-5}. The other four strategies attain accuracies of around 10^{-1} to 10^{-2} for $m = 20$.

Eigenspace recovery for the Random Gaussian strategy. Finally, we focus on the Random Gaussian sampling strategy, which is clearly one of the most successful strategies. In Figure 5.5, we plot in the first row the high-energy modes $\{\phi_{1,2,3,4}\}$ for the reference solution, and in the second row we plot the high-energy modes $\{\phi_{1,2,3,4}^r\}$ obtained from the Random Gaussian strategy with 20 samples. The similarity is evident.

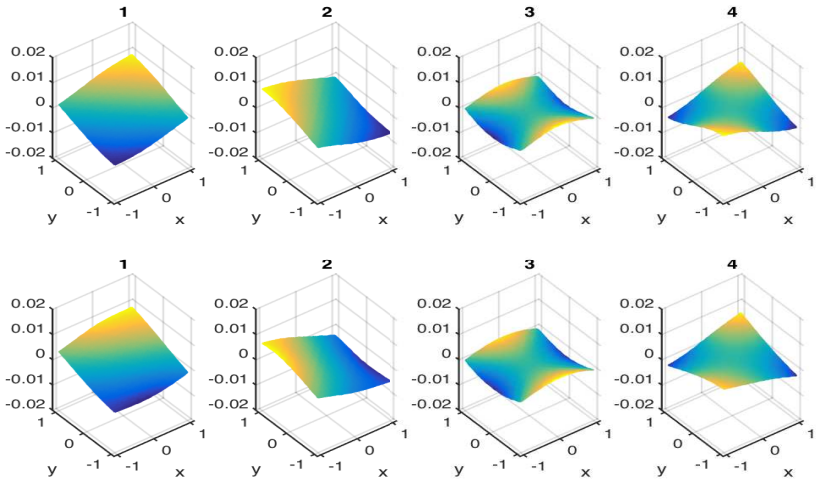

FIG. 5.5. *Recovery of high-energy modes using the Random Gaussian sampling strategy, with* 20 *samples. The first row shows the first four modes of the reference solution; the second row shows the first four modes obtained from the Random Gaussian sampling strategy.*

6. Conclusion. In this paper we study random sampling methods that approximate the optimal solution space that attains Kolmogorov n-width in the context of generalized finite element methods. It is shown that certain random sampling methods capture the main part of the local solution spaces with high accuracy, and that efficiency can be evaluated by the energy contained in the proposed random space.

Numerical comparisons of seven different sampling strategies show that two strategies are superior: Random Gaussian sampling and Smooth Boundary sampling.

Appendix A. Calculation of the Kolmogorov distance. Suppose we are given the optimal space

$$\Phi_m = \operatorname{Span} \mathsf{X}, \quad \text{where } \mathsf{X} = [\phi_1, \dots, \phi_m],$$

and a proposed space

$$\Phi_n^r = \operatorname{Span} \mathsf{Y}, \quad \text{where } \mathsf{Y} = [\phi_1^r, \dots, \phi_n^r], \quad \text{where } n \geq m,$$

such that

$$\langle \phi_i, \phi_j \rangle_{\mathcal{E}(\omega^*)} = \delta_{ij} \quad \text{and} \quad \langle \phi_i^r, \phi_j^r \rangle_{\mathcal{E}(\omega^*)} = \delta_{ij}.$$

Recall the following definition of Kolmogorov distance from (4.10):

$$d(\Phi_n^r, \Phi_m) = \max_{\substack{x \in \Phi_m, \\ \|x\|_{\mathcal{E}(\omega^*)} \leq 1}} \min_{y \in \Phi_n^r} \|x - y\|_{\mathcal{E}(\omega)}.$$

To calculate $d(\Phi_n^r, \Phi_m)$ explicitly, we write $x = \mathsf{X}\alpha$ for some $\alpha \in \mathbb{R}^m$, and $y = \mathsf{Y}\beta$ for some $\beta \in \mathbb{R}^n$. The Kolmogorov distance is achieved when $\|x\|_{\mathcal{E}(\omega^*)} = 1$, which implies that $\|\alpha\| = 1$, where this $\|\cdot\|$ is the usual Euclidean norm on $\mathbb{R}^m$. By expanding the objective, we have

$$\begin{aligned} \frac{1}{2}\|x - y\|_{\mathcal{E}(\omega)}^2 &= \frac{1}{2}\langle \mathsf{X}\alpha - \mathsf{Y}\beta, \mathsf{X}\alpha - \mathsf{Y}\beta \rangle_{\mathcal{E}(\omega)} \\ &= \frac{1}{2}\beta^\top \mathsf{Y}_A \beta - \alpha^\top C_A \beta + \frac{1}{2}\alpha^\top \mathsf{X}_A \alpha, \end{aligned}$$

where $(\mathsf{Y}_A)_{ij} = \langle \phi_i^r, \phi_j^r \rangle_{\mathcal{E}(\omega)}$, $(\mathsf{X}_A)_{ij} = \langle \phi_i, \phi_j \rangle_{\mathcal{E}(\omega)}$, and $(C_A)_{ij} = \langle \phi_j, \phi_i^r \rangle_{\mathcal{E}(\omega)}$. The minimizing value of β is given explicitly by

$$\beta = \mathsf{Y}_A^{-1} C_A \alpha \,,$$

for which we have

$$\frac{1}{2} d(\Phi_n^r, \Phi_m)^2 = \max_{\|\alpha\|=1} \alpha^\top (\mathsf{X}_A - C_A^\top \mathsf{Y}_A^{-1} C_A) \alpha = \|\mathsf{X}_A - C_A^\top \mathsf{Y}_A^{-1} C_A\|_2^2 \,.$$

The Kolmogorov distance is therefore

$$d(\Phi_n^r, \Phi_m) = \sqrt{2} \|\mathsf{X}_A - C_A^\top \mathsf{Y}_A^{-1} C_A\|_2 \,.$$

Appendix B. Well-posedness of energy $E(\mathcal{Z})$. We show that the energy defined in Definition 4.1 is a well-defined quantity. More specifically, given a k-dimensional space $\mathcal{Z}$ and two different B-orthonormal matrices $\mathsf{Z}_1, \mathsf{Z}_2 \in \mathbb{R}^{n \times k}$ whose columns span the space $\mathcal{Z}$, we show that they yield the same value of $E(\mathcal{Z})$:

$$E(\mathcal{Z}) = \frac{\mathrm{Tr}(\mathsf{Z}_1^\top \mathsf{A} \mathsf{Z}_1)}{\mathrm{Tr}(\mathsf{Z}_1^\top \mathsf{B} \mathsf{Z}_1)} = \frac{\mathrm{Tr}(\mathsf{Z}_2^\top \mathsf{A} \mathsf{Z}_2)}{\mathrm{Tr}(\mathsf{Z}_2^\top \mathsf{B} \mathsf{Z}_2)} \,. \tag{B.1}$$

Proof. Since Z_1 and Z_2 share the column space $\mathcal{Z}$, there must exist an invertible matrix $\mathsf{P} \in \mathbb{R}^{k \times k}$ such that $\mathsf{Z}_1 = \mathsf{Z}_2 \mathsf{P}$.

We show first that P is unitary. Because both Z_1 and Z have B-orthonormal columns, we have

$$\mathsf{Z}_1^\top \mathsf{B} \mathsf{Z}_1 = \mathsf{Z}_2^\top \mathsf{B} \mathsf{Z}_2 = \mathsf{I} \,,$$

which implies that

$$\mathsf{I} = \mathsf{Z}_1^\top \mathsf{B} \mathsf{Z}_1 = (\mathsf{Z}_2 \mathsf{P})^\top \mathsf{B} (\mathsf{Z}_2 \mathsf{P}) = \mathsf{P}^\top \mathsf{Z}_2^\top \mathsf{B} \mathsf{Z}_2 \mathsf{P} = \mathsf{P}^\top \mathsf{P}.$$

By definition of Z_1 and Z_2, we have

$$\frac{\mathrm{Tr}(\mathsf{Z}_1^\top \mathsf{A} \mathsf{Z}_1)}{\mathrm{Tr}(\mathsf{Z}_1^\top \mathsf{B} \mathsf{Z}_1)} = \frac{\mathrm{Tr}(\mathsf{Z}_1^\top \mathsf{A} \mathsf{Z}_1)}{k}, \quad \frac{\mathrm{Tr}(\mathsf{Z}_2^\top \mathsf{A} \mathsf{Z}_2)}{\mathrm{Tr}(\mathsf{Z}_2^\top \mathsf{B} \mathsf{Z}_2)} = \frac{\mathrm{Tr}(\mathsf{Z}_2^\top \mathsf{A} \mathsf{Z}_2)}{k},$$

so our claim (B.1) will hold if we can show that $\mathrm{Tr}(\mathsf{Z}_1^\top \mathsf{A} \mathsf{Z}_1) = \mathrm{Tr}(\mathsf{Z}_2^\top \mathsf{A} \mathsf{Z}_2)$. Indeed this follows from

$$\mathrm{Tr}(\mathsf{Z}_1^\top \mathsf{A} \mathsf{Z}_1) = \mathrm{Tr}\left((\mathsf{Z}_2 \mathsf{P})^\top \mathsf{A} (\mathsf{Z}_2 \mathsf{P})\right) = \mathrm{Tr}\left(\mathsf{P}^\top \mathsf{Z}_2^\top \mathsf{A} \mathsf{Z}_2 \mathsf{P}\right) = \mathrm{Tr}\left(\mathsf{Z}_2^\top \mathsf{A} \mathsf{Z}_2 \mathsf{P} \mathsf{P}^\top\right) = \mathrm{Tr}(\mathsf{Z}_2^\top \mathsf{A} \mathsf{Z}_2),$$

where the last equality comes from the orthonormality of P. ∎

Acknowledgment. We are grateful to the two referees for their constructive comments that improved the paper considerably.

REFERENCES

[1] I. Babuška and R. Lipton, *Optimal local approximation spaces for generalized finite element methods with application to multiscale problems*, Multiscale Model. Simul., 9 (2011), pp. 373–406, https://doi.org/10.1137/100791051.

[2] I. Babuška and J. M. Melenk, *The partition of unity method*, Internat. J. Numer. Methods Engrg., 40 (1997), pp. 727–758.

[3] M. Bebendorf, *Why finite element discretizations can be factored by triangular hierarchical matrices*, SIAM J. Numer. Anal., 45 (2007), pp. 1472–1494, https://doi.org/10.1137/060669747.
[4] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, AMS Chelsea Publishing, Providence, RI, 2011.
[5] V. M. Calo, Y. Efendiev, J. Galvis, and G. Li, *Randomized oversampling for generalized multiscale finite element methods*, Multiscale Model. Simul., 14 (2016), pp. 482–501, https://doi.org/10.1137/140988826.
[6] W. E and B. Engquist, *The heterogeneous multi-scale methods*, Commun. Math. Sci., 1 (2003), pp. 87–133.
[7] W. E, P. Ming, and P. Zhang, *Analysis of the heterogeneous multiscale method for elliptic homogenization problems*, J. Amer. Math. Soc., 18 (2005), pp. 121–156.
[8] Y. R. Efendiev, T. Y. Hou, and X.-H. Wu, *Convergence of a nonconforming multiscale finite element method*, SIAM J. Numer. Anal., 37 (2000), pp. 888–910, https://doi.org/10.1137/S0036142997330329.
[9] J. Galvis and Y. Efendiev, *Domain decomposition preconditioners for multiscale flows in high contrast media: Reduced dimension coarse spaces*, Multiscale Model. Simul., 8 (2010), pp. 1621–1644, https://doi.org/10.1137/100790112.
[10] W. Hackbusch, *Hierarchical Matrices: Algorithms and Analysis*, Springer Ser. Comput. Math. 49, Springer, 2015.
[11] N. Halko, P. G. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288, https://doi.org/10.1137/090771806.
[12] T. Y. Hou and X.-H. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.
[13] T. Y. Hou and X.-H. Wu, *A multiscale finite element method for PDEs with oscillatory coefficients*, in Numerical Treatment of Multi-scale Problems (Kiel, 1997), Notes Numer. Fluid Mech. 70, Friedr. Vieweg & Sohn, Braunschweig, 1999, pp. 58–69.
[14] T. Y. Hou, X.-H. Wu, and Z. Cai, *Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients*, Math. Comp., 68 (1999), pp. 913–943.
[15] T. Y. Hou and P. Zhang, *Sparse operator compression of higher-order elliptic operators with rough coefficients*, Res. Math. Sci., 4 (2017), 24.
[16] W. B. Johnson and J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space*, Contemp. Math., 26 (1984), pp. 189–206.
[17] L. Lin, J. Lu, L. Ying, and W. E, *Adaptive local basis set for Kohn–Sham density functional theory in a discontinuous Galerkin framework* I: *Total energy calculation*, J. Comput. Phys., 231 (2012), pp. 2140–2154.
[18] R. Lipton, P. Sinz, and M. Stuebner, *Uncertain loading and quantifying maximum energy concentration within composite structures*, J. Comput. Phys., 325 (2016), pp. 38–52.
[19] A. Målqvist and D. Peterseim, *Localization of elliptic multiscale problems*, Math. Comp., 83 (2014), pp. 2583–2603.
[20] P. Ming and X. Yue, *Numerical methods for multiscale elliptic problems*, J. Comput. Phys., 214 (2006), pp. 421–445.
[21] H. Owhadi, *Bayesian numerical homogenization*, Multiscale Model. Simul., 13 (2015), pp. 812–828, https://doi.org/10.1137/140974596.
[22] H. Owhadi and L. Zhang, *Metric-based upscaling*, Comm. Pure Appl. Math., 60 (2007), pp. 675–723.
[23] H. Owhadi, L. Zhang, and L. Berlyand, *Polyharmonic homogenization, rough polyharmonic splines and sparse super-localization*, ESAIM Math. Model. Numer. Anal., 48 (2014), pp. 517–552.
[24] G. Papanicolaou, *Asymptotic analysis of transport processes*, Bull. Amer. Math. Soc., 81 (1975), pp. 330–392.
[25] A. Pinkus, *n-Widths in Approximation Theory*, Ergeb. Math. Grenzgeb. (3) 7, Springer-Verlag, Berlin, 1985.
[26] M. A. Schweitzer and S. Wu, *Evaluation of local multiscale approximation spaces for partition of unity methods*, in Meshfree Methods for Partial Differential Equations VIII, M. Griebel and M. A. Schweitzer, eds., Springer International, Cham, 2017, pp. 167–198.